Title
Multiround private information retrieval: Capacity and storage overhead

Journal
IEEE Transactions on Information Theory, 64(8)

ISSN
0018-9448

Authors
Sun, H
Jafar, SA

Publication Date
2018-08-01

DOI
10.1109/TIT.2018.2789426

Multiround Private Information Retrieval: Capacity and Storage Overhead

Hua Sun and Syed A. Jafar

Abstract

The capacity has recently been characterized for the private information retrieval (PIR) problem as well as several of its variants. In every case it is assumed that all the queries are generated by the user simultaneously. Here we consider multiround PIR, where the queries in each round are allowed to depend on the answers received in previous rounds. We show that the capacity of multiround PIR is the same as the capacity of single-round PIR (the result is generalized to also include T-privacy constraints). Combined with previous results, this shows that there is no capacity advantage from multiround over single-round schemes, non-linear over linear schemes or from ϵ-error over zero-error schemes. However, we show through an example that there is an advantage in terms of storage overhead. We provide an example of a multiround, non-linear, ϵ-error PIR scheme that requires a strictly smaller storage overhead than the best possible with single-round, linear, zero-error PIR schemes.

Hua Sun (email: huas2@uci.edu) and Syed A. Jafar (email: syed@uci.edu) are with the Center of Pervasive Communications and Computing (CPCC) in the Department of Electrical Engineering and Computer Science (EECS) at the University of California Irvine.

1 Introduction

Private information retrieval (PIR) [1, 2] is one of the canonical problems in theoretical computer science and cryptography. The PIR setting involves K messages that are assumed to be independent, N distributed databases that are replicated (each database stores all K messages) and non-colluding (the databases do not communicate with each other), and a user who desires one of the K messages. A PIR scheme is any mechanism by which a user may retrieve his desired message from the databases privately, i.e., without revealing any information about which message is being retrieved, to any individual database. An information theoretic formulation of PIR guarantees the user's privacy even if the databases are computationally unbounded.[1] The "rate" of a PIR scheme is defined as the ratio of the number of bits of desired information to the total number of bits downloaded by the user from all the databases. The supremum of achievable rates is defined to be the capacity of PIR. For K messages and N databases, the capacity of PIR was characterized recently in [6] as

$$C = \left(1 + 1/N + 1/N^2 + \cdots + 1/N^{K-1}\right)^{-1} \tag{1}$$

The capacity has also been determined for various constrained forms of PIR such as LPIR [7] – where message *lengths* can be arbitrary, TPIR [8] – where any set of up to T databases may collude, RPIR [8] – where *robustness* is required against unresponsive databases, SPIR [9] – which extends the privacy constraint *symmetrically* to protect both the user and the databases, MDS-PIR [10] and MDS-SPIR [11] – variants of PIR and SPIR, respectively, where each message is separately MDS coded.[2]

A common theme in these results is that there is no capacity advantage of non-linear schemes over linear schemes, or of ϵ-error schemes over zero-error schemes. This is a matter of some curiosity because the necessity of non-linear coding schemes has often been a key obstacle in network coding capacity problems [13, 14, 15, 16], and the capacity benefit of ϵ-error schemes over zero-error schemes for network coding problems in general [17] remains one of the key unresolved mysteries — with direct connections to the edge-removal question [18] and the existence of strong converses [19] in network information theory. Motivated by this curiosity, in this work we explore another important variant of PIR – *multiround* PIR (MPIR). Our contributions are summarized next.

The classical PIR setting assumes that all the queries are simultaneously generated by the user. This assumption is also made in [6]. However, such a constraint is not essential to PIR. What if this constraint is relaxed, i.e., multiple rounds of queries and answers are allowed, such that the queries in each round of communication are generated by the user with the knowledge of the answers from all previous rounds? The resulting variant of the PIR problem is the *multiround* PIR (MPIR) problem (also known as interactive PIR [20, 21]). Multiround PIR has been noted as an intriguing possibility in several prior works [2, 20, 21]. However, it is not known whether there is any benefit of MPIR over single-round PIR. Answering this question from a capacity perspective is the first contribution of this work. Specifically, we show that the capacity of MPIR is the same as the capacity of PIR, i.e., both are given by (1). Combined with previous results, this shows that

[1]There is also a widely studied cryptographic formulation of PIR, where the user's privacy is guaranteed only against computationally bounded databases [3, 4, 5].

[2]As a caveat, we note that separate MDS coding of each message is a restrictive assumption. Consider the setting with $K = 2$ messages, $N = 3$ databases and the storage size of each database is equal to the size of one message. If separate MDS codes are employed for each message, then the maximum rate (capacity) is equal to 3/5 [10]. However, Example 2 in [12] shows that rate 2/3 (> 3/5) is achievable with a storage code that jointly encodes both messages.

there is no capacity advantage from multiround over single-round schemes, non-linear over linear schemes or from ϵ-error over zero-error schemes. Furthermore, we show that this is true even with T-privacy constraints.

To complement the capacity analysis, we consider another metric of interest – storage overhead. Classical PIR assumes replicated databases, i.e., each database stores all the messages. For larger datasets, replication schemes incur substantial storage costs. Coding has been shown to be an effective way to reduce the storage costs in distributed data storage systems. Applications of coding to reduce the storage overhead for PIR have attracted attention recently [22, 12, 23, 24, 25, 26, 27, 10, 28, 11]. In this context, our main contribution is an example ($N = 2$ databases, $K = 2$ messages) of a multiround, non-linear, ϵ-error PIR scheme that achieves a strictly smaller storage overhead than the best possible with a single-round, linear, zero-error scheme. The simplicity of the scheme and the $N = K = 2$ setting makes it an attractive point of reference for future work toward understanding the role of linear versus non-linear schemes, zero-error versus ϵ-error capacity, and single-round versus multiround communications. Interestingly, the scheme reveals that coded storage is useful not only for reducing the storage overhead, but also it has a surprising benefit of enhancing the privacy of PIR.

Notation: For $n_1, n_2 \in \mathbb{Z}, n_1 \leq n_2$, define the notation $[n_1 : n_2]$ as the set $\{n_1, n_1+1, \cdots, n_2\}$, $A(n_1 : n_2)$ as the vector $(A(n_1), A(n_1+1), \cdots, A(n_2))$ and $A_{n_1:n_2}$ as the vector $(A_{n_1}, A_{n_1+1}, \cdots, A_{n_2})$. When $n_1 > n_2$, $[n_1 : n_2]$ is a null set and $A(n_1 : n_2), A_{n_1:n_2}$ are null vectors. For an index set $\mathcal{T} = \{i_1, i_2, \cdots, i_n\}$ such that $i_1 < i_2 < \cdots < i_n$, the notation $A_{\mathcal{T}}$ represents the vector $(A_{i_1}, A_{i_2}, \cdots, A_{i_n})$. The notation $X \sim Y$ is used to indicate that X and Y are identically distributed.

2 Problem Statement

Let us start with a general problem statement that can then be specialized to various settings of interest. Consider K independent messages $W_1, \cdots, W_K$, each comprised of L i.i.d. uniform bits.

$$H(W_1, \cdots, W_K) = H(W_1) + \cdots + H(W_K), \tag{2}$$
$$H(W_1) = \cdots = H(W_K) = L. \tag{3}$$

There are N databases. Let S_n denote the information that is stored at the n^{th} database.

$$H(S_n | W_1, W_2, \cdots, W_K) = 0, \quad \forall n \in [1 : N]. \tag{4}$$

Define the storage overhead α as the ratio of the total amount of storage used by all databases to the total amount of data.

$$\alpha \triangleq \frac{\sum_{n=1}^{N} H(S_n)}{KL}. \tag{5}$$

For replication based schemes, each database stores all K messages, so $S_n = (W_1, W_2, \cdots, W_K)$, $H(S_n) = KL$, and the storage overhead, $\alpha = N$.

A user privately generates θ uniformly from $[1 : K]$ and wishes to retrieve W_θ while keeping θ a secret from each database.

Prior works on capacity of PIR and its variants make certain (implicitly justified) assumptions of deterministic behavior, e.g., that the answers provided by the databases are deterministic functions of queries and messages. Here we will follow, instead, an explicit formulation. We allow randomness in the strategies followed by the user and the databases. This is accomplished by representing the actions of the user and the databases as functions of random variables. Let us use $\mathbb{F}$ to denote a random variable privately generated by the user, whose realization is not available to the databases. Similarly, $\mathbb{G}$ is a random variable that determines the random strategies followed by the databases, and whose realizations are assumed to be known to all the databases and the user without loss of generality. $\mathbb{F}$ and $\mathbb{G}$ take values over the set of all deterministic strategies that the user or the databases can follow, respectively, associating each strategy with a certain probability. $\mathbb{F}$ and $\mathbb{G}$ are generated offline, i.e., before the realizations of the messages or the desired message index are known. Since these random variables are generated a-priori we must have

$$\begin{aligned} & H(\theta, \mathbb{F}, \mathbb{G}, W_1, \cdots, W_K) \\ = \;& H(\theta) + H(\mathbb{F}) + H(\mathbb{G}) + H(W_1) + \cdots + H(W_K) \end{aligned} \tag{6}$$

The multiround PIR scheme proceeds as follows. Suppose $\theta = k$. In order to retrieve $W_k, k \in [1:K]$ privately, the user communicates with the databases over Γ rounds. In the first round, the user privately generates N random queries, $Q_1^{[k]}(1), Q_2^{[k]}(1), \cdots, Q_N^{[k]}(1)$.

$$H(Q_1^{[k]}(1), Q_2^{[k]}(1), \cdots, Q_N^{[k]}(1)|\mathbb{F}) = 0, \quad \forall k \in [1:K] \tag{7}$$

The user sends query $Q_n^{[k]}(1)$ to the n^{th} database, $\forall n \in [1:N]$. Upon receiving $Q_n^{[k]}(1)$, the n^{th} database generates an answering string $A_n^{[k]}(1)$. Without loss of generality, we assume that the answering string is a function of $Q_n^{[k]}(1)$, the stored information S_n, and the random variable $\mathbb{G}$.

$$H(A_n^{[k]}(1)|Q_n^{[k]}(1), S_n, \mathbb{G}) = 0. \tag{8}$$

Each database returns to the user its answer $A_n^{[k]}(1)$.

Proceeding similarly[3] , over the γ^{th} round, $\gamma \in [2:\Gamma]$, the user generates N queries $Q_1^{[k]}(\gamma), \cdots, Q_N^{[k]}(\gamma)$, which are functions of previous queries and answers and $\mathbb{F}$,

$$H(Q_{1:N}^{[k]}(\gamma)|Q_{1:N}^{[k]}(1:\gamma-1), A_{1:N}^{[k]}(1:\gamma-1), \mathbb{F}) = 0 \tag{9}$$

The user sends query $Q_n^{[k]}(\gamma)$ to the n^{th} database, which generates an answer $A_n^{[k]}(\gamma)$ and returns $A_n^{[k]}(\gamma)$ to the user. The answer is a function of all queries received so far, the stored information S_n, and $\mathbb{G}$,

$$H(A_n^{[k]}(\gamma)|Q_n^{[k]}(1:\gamma), S_n, \mathbb{G}) = 0. \tag{10}$$

At the end of Γ rounds, from all the information that is now available to the user $(A_{1:N}^{[k]}(1:\Gamma), Q_{1:N}^{[k]}(1:\Gamma), \mathbb{F})$, the user decodes the desired message W_k according to a decoding rule that is specified by the PIR scheme. Let P_e denote the probability of error achieved with the specified decoding rule.

[3]One might wonder if the setting can be further generalized by allowing sequential queries, i.e., allowing the query to each database to depend not only on the answers received from previous rounds, but also on the answers received from other databases queried previously within the same round. We note that sequential queries are already contained in our multiround framework, e.g., by querying only one database in each round (sending null queries to the remaining databases).

To protect the user's privacy, the K possible values of the desired message index should be indistinguishable from the perspective of any subset $\mathcal{T} \subset [1:N]$ of at most T colluding databases, i.e., the following privacy constraint must be satisfied.

$$\begin{aligned}
[T\text{-Privacy}] \quad (Q_{\mathcal{T}}^{[k]}(1:\Gamma), A_{\mathcal{T}}^{[k]}(1:\Gamma), \mathbb{G}, S_{\mathcal{T}}) \quad &\sim \quad (Q_{\mathcal{T}}^{[k']}(1:\Gamma), A_{\mathcal{T}}^{[k']}(1:\Gamma), \mathbb{G}, S_{\mathcal{T}}) \\
&\quad\ \forall k, k' \in [1:K], \forall \mathcal{T} \subset [1:N], |\mathcal{T}| = T
\end{aligned} \tag{11}$$

The PIR rate characterizes how many bits of desired information are retrieved per downloaded bit and is defined as follows.

$$R = \frac{L}{D} \tag{12}$$

where D is the expected value[4] of the total number of bits downloaded by the user from all the databases over all Γ rounds.

A rate R is said to be ϵ-error achievable if there exists a sequence of PIR schemes, indexed by L, each of rate greater than or equal to R, for which $P_e \to 0$ as $L \to \infty$. Note that for such a sequence of PIR schemes, from Fano's inequality we must have

$$\begin{aligned}
[\text{Correctness}] \quad o(L) \quad &= \quad \frac{1}{L} H(W_k | A_{1:N}^{[k]}(1:\Gamma), Q_{1:N}^{[k]}(1:\Gamma), \mathbb{F}) \\
&\overset{(7)(9)}{=} \quad \frac{1}{L} H(W_k | A_{1:N}^{[k]}(1:\Gamma), \mathbb{F}), \quad \forall k \in [1:K]
\end{aligned} \tag{13}$$

where $o(L)$ represents any term whose value approaches zero as L approaches infinity. The supremum of ϵ-error achievable rates is called the ϵ-error capacity C_ϵ.

A rate R is said to be zero-error achievable if there exists (for some L) a PIR scheme of rate greater than or equal to R for which $P_e = 0$. The supremum of zero-error achievable rates is called the zero-error capacity C_o. From the definitions, it is evident that

$$C_o \leq C_\epsilon \tag{14}$$

3 Results

There are two main contributions in this work, summarized in the following sections.

3.1 Capacity Perspective

We first consider the capacity benefits of multiple rounds of communication in the classical setting where each database stores all messages, i.e., storage is unconstrained. We present our result in the general context of multiround PIR with T-privacy constraints (MTPIR). The MTPIR setting is obtained from the general problem statement by relaxing the storage overhead constraints, i.e.,

$$S_n \quad = \quad (W_1, W_2, \cdots, W_K), \forall n \in [1:N]$$

[4]Alternatively, D may be defined as the maximum download needed by the PIR scheme which (similar to choosing zero-error instead of ϵ-error) weakens the converse and strengthens the achievability arguments in general. The capacity characterizations in this work, as well as previous works in [6, 8, 9] hold under either definition. This is because in every case, the upper bounds allow average download D, while the achievability only requires maximum download D.

$$\alpha = N$$

i.e., each database stores all the messages (replication). The following theorem presents the main result.

Theorem 1 *The capacity of MTPIR*

$$C_o = C_\epsilon = \left(1 + T/N + T^2/N^2 + \cdots + T^{K-1}/N^{K-1}\right)^{-1}.$$

The converse proof of Theorem 1 is presented in Section 4. Achievability follows directly from [8]. The following observations place the result in perspective.

1. The capacity of MTPIR matches the capacity of TPIR found in [8], i.e., multiple rounds do not increase capacity.

2. Setting $T = 1$ gives us the capacity of multiround PIR (MPIR) without T-privacy constraints. The capacity of MPIR matches the capacity of PIR found in [6], i.e., multiple rounds do not increase capacity.

3. Since the achievability proofs in [8, 6] only require linear and zero-error schemes, there is no capacity benefit of multiple rounds over single-round schemes, non-linear over linear schemes, or ϵ-error over zero-error schemes.

4. For all N, K, T, Γ the converse proof of Theorem 1 generalizes the converse proofs of [8, 6]. Remarkably, it requires only Shannon information inequalities, i.e., sub-modularity of entropy.

3.2 Storage Overhead Perspective

As summarized above, our first result shows that in a broad sense – with or without colluding databases – there is no capacity benefit of multiple rounds over single-round communication, ϵ-error over zero-error schemes or non-linear over linear schemes for PIR. This pessimistic finding may lead one to believe that there is little reason to further explore interactive communication, non-linear schemes or ϵ-error schemes for PIR. As our main contribution in this section, we offer an optimistic counterpoint by looking at the PIR problem from the perspective of storage overhead instead of capacity. The counterpoint is made through a counterexample. The counterexample is quite remarkable in itself as it shows from a storage overhead perspective not only the advantage of a multiround PIR scheme over all single-round PIR schemes, but also of a non-linear PIR scheme over all linear PIR schemes, and an ϵ-error scheme over all zero-error schemes.

For a counterexample the simplest setting is typically the most interesting. Therefore, in this section we will only consider the simplest non-trivial setting, with $K = 2$ messages, $N = 2$ databases, and $T = 1$, i.e., no collusion among databases. Recall that for this setting the capacity is $C = 2/3$. For our counterexample we explore the minimum storage overhead that is needed to achieve the rate $2/3$.

Theorem 2 *For $K = 2, N = 2, T = 1$, and for rate $2/3$,*

1. *there exists a multiround, non-linear and ϵ-error PIR scheme with storage overhead*

$$\alpha = 3/4 + 3/8 \log_2 3$$

which is less than $3/2$.

2. the storage overhead of any single-round, linear and zero-error PIR scheme is

$$\alpha \geq 3/2$$

The achievability arguments, including the multiround, non-linear and ϵ-error PIR scheme that proves the first part of Theorem 2 are presented in this section. The proof of the second claim notably utilizes Ingleton's inequality, which goes beyond submodularity, and is presented in Section 5.

3.2.1 A multiround, non-linear and ϵ-error PIR scheme for $K = 2, N = 2, T = 1$

Define w_1, w_2 as two independent uniform binary random variables. Further, define

$$\begin{aligned} x_1 &= w_1 \wedge w_2 && (15)\\ x_2 &= (\sim w_1) \wedge (\sim w_2) && (16)\\ y_1 &= w_1 \wedge (\sim w_2) && (17)\\ y_2 &= (\sim w_1) \wedge w_2 && (18) \end{aligned}$$

where $\wedge$ and $\sim$ are the logical AND and NOT operators. Note the following,

$$\begin{aligned} x_1 = 1 &\Rightarrow (w_1, w_2) = (1, 1) && (19)\\ x_2 = 1 &\Rightarrow (w_1, w_2) = (0, 0) && (20)\\ x_1 = 0 &\Rightarrow (w_1, w_2) = (y_1, y_2) && (21)\\ x_2 = 0 &\Rightarrow (w_1, w_2) = (\sim y_2, \sim y_1) && (22) \end{aligned}$$

For ease of exposition, consider first the case where each message is only one bit long. In this case, the messages W_1, W_2, directly correspond to w_1, w_2, respectively. Denote the first database as DB1 and the second database as DB2. Regardless of whether the user desires W_1 or W_2, he flips a private fair coin, and requests the value of either x_1 or x_2 from DB1. If the answer is 1, then according to (19) and (20) the user knows the values of both w_1, w_2 and no further information is requested from DB2. If the answer is 0, then the user proceeds as follows.

- If $x_1 = 0$ and W_1 is desired, ask DB2 for the value of y_1. Retrieve $w_1 = y_1$.
- If $x_1 = 0$ and W_2 is desired, ask DB2 for the value of y_2. Retrieve $w_2 = y_2$.
- If $x_2 = 0$ and W_1 is desired, ask DB2 for the value of y_2. Retrieve $w_1 =\sim y_2$.
- If $x_2 = 0$ and W_2 is desired, ask DB2 for the value of y_1. Retrieve $w_2 =\sim y_1$.

Note that in order to answer the user's queries, DB1 only needs to store (x_1, x_2), and DB2 only needs to store (y_1, y_2). This observation is the key to not only the reduced storage overhead, but also the enhanced privacy of this scheme.

Further, in preparation for the proofs that follow, let us define another binary random variable u, which takes the value $u = 0$ if no response is needed from DB2, and the value $u = 1$ otherwise. Note that $u = 0$ implies that $(y_1, y_2) = (0, 0)$. On the other hand, if $u = 1$, then (y_1, y_2) takes the values $(0, 0), (1, 0), (0, 1)$, each with probability 1/3. Therefore,

$$H(y_1, y_2|u) = 1/4 \times H(y_1, y_2|u = 0) + 3/4 \times H(y_1, y_2|u = 1) \quad (23)$$

$$= \quad 1/4 \times 0 + 3/4 \times H(1/3, 1/3, 1/3) = 3/4 \log_2 3 \tag{24}$$

The correctness of the scheme is obvious from (19)-(22). Let us verify that the scheme is private. Start with DB1. The query to DB1 is equally likely to be x_1 or x_2, regardless of the desired message index and the message realizations. Therefore, DB1 learns nothing about which message is retrieved. Next consider DB2. Let us prove that $(Q_2^{[1]}, y_1, y_2) \sim (Q_2^{[2]}, y_1, y_2)$.

$$\begin{array}{c|c} \multicolumn{2}{c}{(\theta=1)} \\ (Q_2^{[1]}, y_1, y_2) & \text{Prob.} \\ \hline (\emptyset, 0, 0) & 1/4 \\ (\text{``}y_1\text{''}, 0, 0) & 1/8 \\ (\text{``}y_2\text{''}, 0, 0) & 1/8 \\ (\text{``}y_1\text{''}, 0, 1) & 1/8 \\ (\text{``}y_2\text{''}, 0, 1) & 1/8 \\ (\text{``}y_1\text{''}, 1, 0) & 1/8 \\ (\text{``}y_2\text{''}, 1, 0) & 1/8 \end{array} \quad \sim \quad \begin{array}{c|c} \multicolumn{2}{c}{(\theta=2)} \\ (Q_2^{[2]}, y_1, y_2) & \text{Prob.} \\ \hline (\emptyset, 0, 0) & 1/4 \\ (\text{``}y_1\text{''}, 0, 0) & 1/8 \\ (\text{``}y_2\text{''}, 0, 0) & 1/8 \\ (\text{``}y_1\text{''}, 0, 1) & 1/8 \\ (\text{``}y_2\text{''}, 0, 1) & 1/8 \\ (\text{``}y_1\text{''}, 1, 0) & 1/8 \\ (\text{``}y_2\text{''}, 1, 0) & 1/8 \end{array}$$

where the double quote notation around a random variable represents the query about its realization. The computation of the joint distribution values is straightforward. We present the derivation here for one case. All other cases follow similarly. From the law of total probability, we have

$$\begin{aligned} &\Pr\left((Q_2^{[1]}, y_1, y_2) = (\text{``}y_1\text{''}, 0, 1)\right) \\ &= \Pr\left((Q_2^{[1]}, y_1, y_2) = (\text{``}y_1\text{''}, 0, 1) | (Q_1^{[1]}, w_1, w_2) = (\text{``}x_1\text{''}, 0, 1)\right) \times \Pr\left((Q_1^{[1]}, w_1, w_2) = (\text{``}x_1\text{''}, 0, 1)\right) \\ &+ \Pr\left((Q_2^{[1]}, y_1, y_2) = (\text{``}y_1\text{''}, 0, 1) | (Q_1^{[1]}, w_1, w_2) = (\text{``}x_2\text{''}, 0, 1)\right) \times \Pr\left((Q_1^{[1]}, w_1, w_2) = (\text{``}x_2\text{''}, 0, 1)\right) \end{aligned} \tag{25}$$

$$= 1 \times 1/8 + 0 \times 1/8 = 1/8 \tag{26}$$

Similarly,

$$\begin{aligned} &\Pr\left((Q_2^{[2]}, y_1, y_2) = (\text{``}y_1\text{''}, 0, 1)\right) \\ &= \Pr\left((Q_2^{[2]}, y_1, y_2) = (\text{``}y_1\text{''}, 0, 1) | (Q_1^{[2]}, w_1, w_2) = (\text{``}x_1\text{''}, 0, 1)\right) \times \Pr\left((Q_1^{[2]}, w_1, w_2) = (\text{``}x_1\text{''}, 0, 1)\right) \\ &+ \Pr\left((Q_2^{[2]}, y_1, y_2) = (\text{``}y_1\text{''}, 0, 1) | (Q_1^{[2]}, w_1, w_2) = (\text{``}x_2\text{''}, 0, 1)\right) \times \Pr\left((Q_1^{[2]}, w_1, w_2) = (\text{``}x_2\text{''}, 0, 1)\right) \end{aligned} \tag{27}$$

$$= 0 \times 1/8 + 1 \times 1/8 = 1/8 \tag{28}$$

Thus, $\Pr\left((Q_2^{[1]}, y_1, y_2) = (\text{``}y_1\text{''}, 0, 1)\right) = \Pr\left((Q_2^{[1]}, y_1, y_2) = (\text{``}y_1\text{''}, 0, 1)\right)$. All other cases are verified similarly. Then, since the distribution of $(Q_2^{[\theta]}, y_1, y_2)$ does not depend on θ, and the answers are only deterministic functions of the query and the stored information, it follows that the scheme is private.

Next consider the L length extension of this PIR scheme, where each desired bit is retrieved independently as described above. Under the L length extension, $W_1, W_2, X_1, X_2, Y_1, Y_2, U$ are sequences of length L, such that the sequence of tuples $[(W_1(l), W_2(l),\ X_1(l), X_2(l),\ Y_1(l), Y_2(l), U(l))]_{l=1}^{L}$ is i.i.d. $\sim (w_1, w_2, x_1, x_2, y_1, y_2, u)$. Since the extension is obtained by repeated independent applications of the PIR scheme described above for retrieving each message bit, it follows trivially that the extended PIR scheme is also correct and private. The purpose for the L length extension, with $L \to \infty$, is to invoke fundamental limits of data compression which optimize both the data rates and the storage overhead as explained next.

Let us show that the rate $2/3$ is achieved asymptotically as $L \to \infty$. We take advantage of the fact that the answers from the databases are not uniformly distributed, and therefore the sequence of answers from each database is compressible. With optimal compression, the user downloads $H(1/4, 3/4)$ bits per desired message bit from DB1. This is because, for each retrieved bit, the answer from DB1 takes the value 1 with probability $1/4$ and 0 with probability $3/4$. From DB2, we download $1/4 \times 0 + 3/4 \times H(1/3, 2/3) = 3/4H(1/3, 2/3)$ bits per desired message bit, because with probability $1/4$ (when the answer from DB1 is 1), no response is requested from DB2 and otherwise within the remaining space of probability measure $3/4$ (when the answer from DB1 is 0), the answer from DB2 is 1 with conditional probability $1/3$ and 0 with conditional probability $2/3$. Therefore the total download is $H(1/4, 3/4) + 3/4H(1/3, 2/3) = 3/2$ bits per desired message bit and the rate achieved is $2/3$.

Next let us determine the storage requirements of this scheme. DB1 needs (X_1, X_2) to answer the user's queries, so with optimal compression, it needs to store $H(x_1, x_2) = H(1/4, 1/4, 1/2) = 3/2$ bits per desired message bit. One might naively imagine that the same storage requirement also applies to DB2, because DB2 similarly needs the values (Y_1, Y_2) to answer the user's queries. However, this is not true, because the query sent to DB2 already contains some information about the message realizations,[5] and this *side-information* allows DB2 to reduce its storage requirement by taking advantage of Slepian Wolf coding [29, 30] (distributed compression with decoder side information).

The key is to realize that DB2 does not need to know (Y_1, Y_2) until after it receives the query from the user. The query from the user includes U as side information. Therefore, using Slepian Wolf coding, DB2 is able to optimally compress the i.i.d. sequence (Y_1, Y_2) to the conditional entropy $H(y_1, y_2|u)$ bits per desired message bit and still decode the (Y_1, Y_2) sequence when it is needed, i.e., after the query is provided by the user. Thus, the total storage required by this PIR scheme is $3/2 + 3/4 \log_2 3$ bits per bit of desired message. Since there are two messages, the storage overhead is $3/4 + 3/8 \log_2 3$.

The following observations are useful to place the new PIR scheme in perspective.

1. The optimal compression guarantees are only available in the ϵ-error sense. Therefore, this PIR scheme is essentially an ϵ-error scheme.

2. The multiround scheme is in fact a sequential PIR scheme that utilizes only one round of queries for each database (two rounds total since there are two databases).

3. The scheme is essentially non-linear because, e.g., the logical AND operator is non-linear.

4. Since the multiround, non-linear and ϵ-error aspects are all essential for *this* scheme to create an advantage in terms of storage overhead, it is an intriguing question whether all three aspects are necessary in *general* or if it is possible to achieve storage overhead less than $3/2$ through another scheme while sacrificing at least one of the three aspects.

[5] Note that the query sent to DB2 is independent of the desired message index but not the message realizations.

5. A key insight from this PIR scheme is the surprising privacy benefit of storage overhead optimization. By not storing all the information at each database, and by optimally compressing the stored information, not only do we reduce the storage overhead, but also we enable stronger privacy guarantees than would hold otherwise. Note that if each database stores all the information (both W_1 and W_2), then the scheme is not private. To see this, suppose $(w_1, w_2) = (1, 1)$. This would be known to DB2 because it stores both messages. Under this circumstance, DB2 knows that if the user asks for y_2, then his desired message must be W_1 and if the user asks for y_1 then his desired message must be W_2. Thus, storing all the information at each database would result in loss of privacy. As another example, we note that if the data is not in its optimally compressed form, i.e., w_1 and/or w_2 are not uniformly distributed then again the PIR scheme would lose privacy. To see this, suppose $\Pr(w_1 = 1)=\Pr(w_2 = 1) > 1/2$. Then DB2 is more likely to be asked for y_1 if the desired message is W_2 than if the desired message is W_1. On the other hand, note that optimal data compression is a pre-requisite in any case for the optimization of rate and storage overhead.[6]

6. Let us consider momentarily the restricted message size setting, where each message is only $L = 1$ bit long. Then it is easy to see that any single-round scheme (all queries generated simultaneously) must download at least 2 bits on average, but our multiround scheme requires an expected download of only $1+3/4 = 7/4$ bits. Thus, even though the download advantage of multiround PIR disappears under unconstrained message lengths, for constrained message lengths there are benefits of multiround PIR.

3.2.2 A single-round, linear and zero-error scheme for $K = 2, N = 2, T = 1$

For comparison, the corresponding scheme from [6] which also achieves rate 2/3 is reproduced below. This will be shown to be the optimal single-round, linear, zero-error scheme for storage overhead in Section 5. Denote the messages, each comprised of 4 bits, as $W_1 = (a_1, a_2, a_3, a_4), W_2 = (b_1, b_2, b_3, b_4)$. The downloaded information from each database is shown below.

	Prob. 1/2		Prob. 1/2	
	Want W_1	Want W_2	Want W_1	Want W_2
Database 1	$a_1, b_1, a_2 + b_2$	$a_1, b_1, a_2 + b_2$	$a_3, b_3, a_4 + b_4$	$a_3, b_3, a_4 + b_4$
Database 2	$a_4, b_2, a_3 + b_1$	$a_2, b_4, a_1 + b_3$	$a_2, b_4, a_1 + b_3$	$a_4, b_2, a_3 + b_1$

The scheme achieves rate 2/3 and is linear, single-round, and zero-error. A total of 6 bits are stored at each database

$$S_1 = (a_1, a_3, b_1, b_3, a_2 + b_2, a_4 + b_4) \tag{29}$$

$$S_2 = (a_2, a_4, b_2, b_4, a_3 + b_1, a_1 + b_3) \tag{30}$$

Thus, the storage overhead is 3/2.

4 Proof of Theorem 1

We first present two useful lemmas. Note that in the proofs, the relevant equations needed to justify each step are specified by the equation numbers set on top of the (in)equality symbols.

[6] Since optimal compression limits are typically achieved asymptotically, if the data is not assumed to be uniform a-priori, then as noted by [20, 21] the privacy guarantees would also be subject to ϵ-leakage that approaches zero as message length approaches infinity.

Lemma 1 *For all $k \in [2:K]$,*

$$\begin{aligned} & I(W_{k:K}; Q_{1:N}^{[k-1]}(1:\Gamma), A_{1:N}^{[k-1]}(1:\Gamma), \mathbb{F}|W_{1:k-1}, \mathbb{G}) \\ \geq \quad & \frac{T}{N} I(W_{k+1:K}; Q_{1:N}^{[k]}(1:\Gamma), A_{1:N}^{[k]}(1:\Gamma), \mathbb{F}|W_{1:k}, \mathbb{G}) + \frac{LT}{N}(1 - o(L)). \end{aligned} \tag{31}$$

Proof:

$$NI(W_{k:K}; Q_{1:N}^{[k-1]}(1:\Gamma), A_{1:N}^{[k-1]}(1:\Gamma), \mathbb{F}|W_{1:k-1}, \mathbb{G})$$

$$\geq \quad \frac{N}{\binom{N}{T}} \sum_{\mathcal{T} \subset [1:N], |\mathcal{T}|=T} I(W_{k:K}; Q_{\mathcal{T}}^{[k-1]}(1:\Gamma), A_{\mathcal{T}}^{[k-1]}(1:\Gamma)|W_{1:k-1}, \mathbb{G}) \tag{32}$$

$$\overset{(11)}{=} \quad \frac{N}{\binom{N}{T}} \sum_{\mathcal{T} \subset [1:N], |\mathcal{T}|=T} I(W_{k:K}; Q_{\mathcal{T}}^{[k]}(1:\Gamma), A_{\mathcal{T}}^{[k]}(1:\Gamma)|W_{1:k-1}, \mathbb{G}) \tag{33}$$

$$= \quad \frac{N}{\binom{N}{T}} \sum_{\mathcal{T} \subset [1:N], |\mathcal{T}|=T} \sum_{\gamma=1}^{\Gamma} I(W_{k:K}; Q_{\mathcal{T}}^{[k]}(\gamma), A_{\mathcal{T}}^{[k]}(\gamma)|Q_{\mathcal{T}}^{[k]}(1:\gamma-1), A_{\mathcal{T}}^{[k]}(1:\gamma-1), W_{1:k-1}, \mathbb{G})$$

$$\geq \quad \frac{N}{\binom{N}{T}} \sum_{\mathcal{T} \subset [1:N], |\mathcal{T}|=T} \sum_{\gamma=1}^{\Gamma} I(W_{k:K}; A_{\mathcal{T}}^{[k]}(\gamma)|Q_{\mathcal{T}}^{[k]}(1:\gamma), A_{\mathcal{T}}^{[k]}(1:\gamma-1), W_{1:k-1}, \mathbb{G}) \tag{34}$$

$$\overset{(8)(10)}{=} \quad \frac{N}{\binom{N}{T}} \sum_{\mathcal{T} \subset [1:N], |\mathcal{T}|=T} \sum_{\gamma=1}^{\Gamma} H(A_{\mathcal{T}}^{[k]}(\gamma)|Q_{\mathcal{T}}^{[k]}(1:\gamma), A_{\mathcal{T}}^{[k]}(1:\gamma-1), W_{1:k-1}, \mathbb{G}) \tag{35}$$

$$\geq \quad \frac{N}{\binom{N}{T}} \sum_{\mathcal{T} \subset [1:N], |\mathcal{T}|=T} \sum_{\gamma=1}^{\Gamma} H(A_{\mathcal{T}}^{[k]}(\gamma)|Q_{1:N}^{[k]}(1:\gamma), A_{1:N}^{[k]}(1:\gamma-1), W_{1:k-1}, \mathbb{F}, \mathbb{G}) \tag{36}$$

$$\geq \quad T \sum_{\gamma=1}^{\Gamma} H(A_{1:N}^{[k]}(\gamma)|Q_{1:N}^{[k]}(1:\gamma), A_{1:N}^{[k]}(1:\gamma-1), W_{1:k-1}, \mathbb{F}, \mathbb{G}) \quad \text{(Han's inequality [30])} \tag{37}$$

$$\overset{(8)(10)}{=} \quad T \sum_{\gamma=1}^{\Gamma} I(W_{k:K}; A_{1:N}^{[k]}(\gamma)|Q_{1:N}^{[k]}(1:\gamma), A_{1:N}^{[k]}(1:\gamma-1), W_{1:k-1}, \mathbb{F}, \mathbb{G}) \tag{38}$$

$$\overset{(7)(9)}{=} \quad T \sum_{\gamma=1}^{\Gamma} I(W_{k:K}; Q_{1:N}^{[k]}(\gamma), A_{1:N}^{[k]}(\gamma)|Q_{1:N}^{[k]}(1:\gamma-1), A_{1:N}^{[k]}(1:\gamma-1), W_{1:k-1}, \mathbb{F}, \mathbb{G}) \tag{39}$$

$$= \quad TI(W_{k:K}; Q_{1:N}^{[k]}(1:\Gamma), A_{1:N}^{[k]}(1:\Gamma)|W_{1:k-1}, \mathbb{F}, \mathbb{G}) \tag{40}$$

$$\overset{(13)}{=} \quad TI(W_{k:K}; W_k, Q_{1:N}^{[k]}(1:\Gamma), A_{1:N}^{[k]}(1:\Gamma)|W_{1:k-1}, \mathbb{F}, \mathbb{G}) - o(L)LT \tag{41}$$

$$\begin{aligned} = \quad & TI(W_{k:K}; W_k|W_{1:k-1}, \mathbb{F}, \mathbb{G}) - o(L)LT \\ & + TI(W_{k+1:K}; Q_{1:N}^{[k]}(1:\Gamma), A_{1:N}^{[k]}(1:\Gamma)|W_{1:k}, \mathbb{F}, \mathbb{G}) \end{aligned} \tag{42}$$

$$\overset{(6)}{=} \quad LT(1 - o(L)) + TI(W_{k+1:K}; Q_{1:N}^{[k]}(1:\Gamma), A_{1:N}^{[k]}(1:\Gamma)|W_{1:k}, \mathbb{F}, \mathbb{G}) \tag{43}$$

$$\overset{(6)}{=} \quad LT(1 - o(L)) + TI(W_{k+1:K}; Q_{1:N}^{[k]}(1:\Gamma), A_{1:N}^{[k]}(1:\Gamma), \mathbb{F}|W_{1:k}, \mathbb{G}) \tag{44}$$

■

Lemma 2

$$I(W_{2:K}; Q_{1:N}^{[1]}(1:\Gamma), A_{1:N}^{[1]}(1:\Gamma), \mathbb{F}|W_1, \mathbb{G}) \leq L(1/R - 1 + o(L)). \tag{45}$$

Proof:

$$\begin{aligned}
& I(W_{2:K}; Q_{1:N}^{[1]}(1:\Gamma), A_{1:N}^{[1]}(1:\Gamma), \mathbb{F}|W_1, \mathbb{G}) \\
\overset{(6)}{=} \quad & I(W_{2:K}; Q_{1:N}^{[1]}(1:\Gamma), A_{1:N}^{[1]}(1:\Gamma), W_1, \mathbb{F}, \mathbb{G}) && (46) \\
\overset{(7)(9)}{=} \quad & I(W_{2:K}; A_{1:N}^{[1]}(1:\Gamma), W_1, \mathbb{F}, \mathbb{G}) && (47) \\
= \quad & I(W_{2:K}; A_{1:N}^{[1]}(1:\Gamma), \mathbb{F}, \mathbb{G}) + I(W_{2:K}; W_1 | A_{1:N}^{[1]}(1:\Gamma), \mathbb{F}, \mathbb{G}) && (48) \\
\overset{(6)(13)}{=} \quad & I(W_{2:K}; A_{1:N}^{[1]}(1:\Gamma) | \mathbb{F}, \mathbb{G}) + o(L)L && (49) \\
= \quad & H(A_{1:N}^{[1]}(1:\Gamma)|\mathbb{F}, \mathbb{G}) - H(A_{1:N}^{[1]}(1:\Gamma)|\mathbb{F}, \mathbb{G}, W_{2:K}) + o(L)L && (50) \\
\overset{(12)}{\leq} \quad & L/R - H(A_{1:N}^{[1]}(1:\Gamma)|\mathbb{F}, \mathbb{G}, W_{2:K}) + o(L)L && (51) \\
\overset{(13)}{=} \quad & L/R - H(W_1, A_{1:N}^{[1]}(1:\Gamma)|\mathbb{F}, \mathbb{G}, W_{2:K}) + o(L)L && (52) \\
\leq \quad & L/R - H(W_1|\mathbb{F}, \mathbb{G}, W_{2:K}) + o(L)L && (53) \\
\overset{(6)}{=} \quad & L/R - L + o(L)L = L(1/R - 1 + o(L)) && (54)
\end{aligned}$$

■

With Lemma 1 and Lemma 2, we are ready to prove the converse.

Rate Outerbound

Starting from $k = 2$ and applying (31) repeatedly for $k \in [3:K]$,

$$\begin{aligned}
& I(W_{2:K}; Q_{1:N}^{[1]}(1:\Gamma), A_{1:N}^{[1]}(1:\Gamma), \mathbb{F}|W_1, \mathbb{G}) \\
\overset{(31)}{\geq} \quad & \frac{T}{N} I(W_{3:K}; Q_{1:N}^{[2]}(1:\Gamma), A_{1:N}^{[2]}(1:\Gamma), \mathbb{F}|W_1, W_2, \mathbb{G}) + \frac{LT(1-o(L))}{N} \\
\overset{(31)}{\geq} \quad & \cdots && (55) \\
\overset{(31)}{\geq} \quad & \frac{T^{K-2}}{N^{K-2}} I(W_K; Q_{1:N}^{[K-1]}(1:\Gamma), A_{1:N}^{[K-1]}(1:\Gamma), \mathbb{F}|W_{1:K-1}, \mathbb{G}) \\
& + \frac{LT(1-o(L))}{N} + \cdots + \frac{LT^{K-2}(1-o(L))}{N^{K-2}} \\
\overset{(31)}{\geq} \quad & \frac{T^{K-2}}{N^{K-2}} \frac{LT(1-o(L))}{N} + \frac{LT(1-o(L))}{N} + \cdots + \frac{LT^{K-2}(1-o(L))}{N^{K-2}} && (56) \\
= \quad & L(1-o(L))(T/N + \cdots + T^{K-1}/N^{K-1}) && (57)
\end{aligned}$$

Combining (57) and (45), we have

$$L(1/R - 1 + o(L)) \geq L(1-o(L))(T/N + \cdots + T^{K-1}/N^{K-1}) \tag{58}$$

Normalizing by L and letting L go to infinity gives us

$$1/R - 1 \quad \geq \quad T/N + \cdots + T^{K-1}/N^{K-1} \tag{59}$$

$$\Rightarrow R \quad \leq \quad (1 + T/N + \cdots + T^{K-1}/N^{K-1})^{-1} \tag{60}$$

thus, the proof is complete.

5 Proof of Theorem 2 – Statement 2.

We show that when $K = 2, N = 2, T = 1, \Gamma = 1$ and the rate equals 2/3, the storage overhead of all zero-error, linear, and single-round PIR schemes is no less than 3/2. Since we only consider single-round schemes in this section, we will simplify the notation, e.g., instead of $Q_2^{[1]}(1)$ we write simply $Q_2^{[1]}$. In addition, without loss of generality, let us make the following simplifying assumptions.

1. We assume that the PIR scheme is symmetric, in that

$$H(A_1^{[1]}|\mathbb{F},\mathbb{G}) \quad = \quad H(A_2^{[1]}|\mathbb{F},\mathbb{G}) = H(A_2^{[2]}|\mathbb{F},\mathbb{G}) \tag{61}$$

$$H(S_1) \quad = \quad H(S_2) \tag{62}$$

   Given any (asymmetric) PIR scheme that retrieves messages of size L, a symmetric PIR scheme with the same rate and storage overhead that retrieves messages of size NL is obtained by repeating the original scheme N times, and in the n^{th} repetition shifting the database indices cyclically by n. This symmetrization process is described in Theorem 3 (see Section 5.1).

2. We assume that $Q_1^{[1]} = Q_1^{[2]}$, i.e., the query for the first database is chosen without the knowledge of the desired message index. There is no loss of generality in this assumption because of the privacy constraint, which requires that $Q_1^{[\theta]}$ is independent of θ.[7] Note that this also means that $A_1^{[1]} = A_1^{[2]}$.

Our goal is to prove a lower bound on the storage overhead. Since the PIR scheme is symmetric by assumption, the storage overhead is $(H(S_1) + H(S_2))/2L = H(S_2)/L$. Furthermore, $H(S_2) \geq H(A_2^{[1]}, A_2^{[2]}|\mathbb{F},\mathbb{G})$, so we will prove a lower bound on $H(A_2^{[1]}, A_2^{[2]}|\mathbb{F},\mathbb{G})$ instead.

Let us start with a useful lemma that holds for all linear and non-linear schemes.

Lemma 3

$$H(A_1^{[1]}|W_1,\mathbb{F},\mathbb{G}) \quad = \quad H(A_2^{[2]}|W_1,\mathbb{F},\mathbb{G}) = H(A_2^{[2]}|W_2,\mathbb{F},\mathbb{G}) = L/2 \tag{63}$$

$$H(A_2^{[2]}|W_1, A_2^{[1]},\mathbb{F},\mathbb{G}) \quad = \quad H(A_2^{[2]}|W_2, A_2^{[1]},\mathbb{F},\mathbb{G}) = L/2 \tag{64}$$

Proof: We prove (63) first. On the one hand, after substituting[8] $R = 2/3$ in Lemma 2, from (47) we have

$$L/2 \quad \geq \quad I(W_2; A_1^{[1]}, A_2^{[1]}, W_1, \mathbb{F}, \mathbb{G}) \tag{65}$$

$$\overset{(6)}{=} \quad I(W_2; A_1^{[1]}, A_2^{[1]}|W_1, \mathbb{F}, \mathbb{G}) \tag{66}$$

[7] Note that instead of $Q_1^{[1]} = Q_1^{[2]}$, we could equivalently assume that $Q_2^{[1]} = Q_2^{[2]}$ without of loss of generality (because privacy also requires that $Q_2^{[\theta]}$ is independent of θ). However, if we simultaneously assume both $Q_1^{[1]} = Q_1^{[2]}$ *and* $Q_2^{[1]} = Q_2^{[2]}$, then there is a loss of generality because together $(Q_1^{[\theta]}, Q_2^{[\theta]})$ is *not* required to be independent of θ by the privacy constraint.

[8] Since we are considering only zero-error schemes, the $o(L)$ term in Lemma 2 is exactly 0.

$$\stackrel{(7)(8)(4)}{=} \quad H(A_1^{[1]}, A_2^{[1]} | W_1, \mathbb{F}, \mathbb{G}) \tag{67}$$

$$\Rightarrow L/2 \quad \geq \quad H(A_1^{[1]} | W_1, \mathbb{F}, \mathbb{G}) \tag{68}$$

$$\text{and } L/2 \quad \geq \quad H(A_2^{[1]} | W_1, \mathbb{F}, \mathbb{G}) \tag{69}$$

On the other hand, from (32) in Lemma 1, we have

$$
\begin{aligned}
L &\leq I(W_2; Q_1^{[1]}, A_1^{[1]} | W_1, \mathbb{G}) + I(W_2; Q_2^{[1]}, A_2^{[1]} | W_1, \mathbb{G}) && (70)\\
&\leq I(W_2; Q_1^{[1]}, A_1^{[1]}, \mathbb{F} | W_1, \mathbb{G}) + I(W_2; Q_2^{[1]}, A_2^{[1]}, \mathbb{F} | W_1, \mathbb{G}) && (71)\\
&\stackrel{(6)}{=} I(W_2; Q_1^{[1]}, A_1^{[1]} | W_1, \mathbb{F}, \mathbb{G}) + I(W_2; Q_2^{[1]}, A_2^{[1]} | W_1, \mathbb{F}, \mathbb{G}) && (72)\\
&\stackrel{(7)(8)(4)}{=} H(A_1^{[1]} | W_1, \mathbb{F}, \mathbb{G}) + H(A_2^{[1]} | W_1, \mathbb{F}, \mathbb{G}) && (73)
\end{aligned}
$$

Combining (68), (69) and (73), we have shown that

$$H(A_1^{[1]} | W_1, \mathbb{F}, \mathbb{G}) = H(A_2^{[1]} | W_1, \mathbb{F}, \mathbb{G}) = L/2 \tag{74}$$

Symmetrically, it follows that $H(A_2^{[2]} | W_2, \mathbb{F}, \mathbb{G}) = L/2$. We are left to prove $H(A_2^{[2]} | W_1, \mathbb{F}, \mathbb{G}) = L/2$. On the one hand, from (68) and (69), we have

$$
\begin{aligned}
L/2 &\geq H(A_1^{[1]} | W_1, \mathbb{F}, \mathbb{G}) = H(A_1^{[2]} | W_1, \mathbb{F}, \mathbb{G}) \quad (\text{Using } A_1^{[1]} = A_1^{[2]}) && (75)\\
L/2 &\geq H(A_2^{[1]} | W_1, \mathbb{F}, \mathbb{G}) && (76)\\
&\stackrel{(7)}{=} H(A_2^{[1]} | W_1, Q_2^{[1]}, \mathbb{F}, \mathbb{G}) && (77)\\
&= H(A_2^{[1]} | W_1, Q_2^{[1]}, \mathbb{G}) && (78)\\
&= H(A_2^{[2]} | W_1, Q_2^{[2]}, \mathbb{G}) && (79)\\
&= H(A_2^{[2]} | W_1, Q_2^{[2]}, \mathbb{F}, \mathbb{G}) && (80)\\
&\stackrel{(7)}{=} H(A_2^{[2]} | W_1, \mathbb{F}, \mathbb{G}) && (81)
\end{aligned}
$$

where (79) follows from the fact that for single-round PIR, the desired message index is independent of the messages, queries and answers, i.e., from (6), we have

$$
\begin{aligned}
& I(\theta; W_1, W_2, \mathbb{F}, \mathbb{G}) = 0 && (82)\\
\stackrel{(7)}{\Longrightarrow}\ & I(\theta; W_1, W_2, \mathbb{F}, \mathbb{G}, Q_2^{[\theta]}) = 0 && (83)\\
\stackrel{(8)(4)}{\Longrightarrow}\ & I(\theta; W_1, W_2, \mathbb{F}, \mathbb{G}, Q_2^{[\theta]}, A_2^{[\theta]}) = 0 && (84)\\
\Longrightarrow\ & A_2^{[1]}, W_1, Q_2^{[1]}, \mathbb{G} \sim A_2^{[2]}, W_1, Q_2^{[2]}, \mathbb{G} && (85)
\end{aligned}
$$

(78) and (80) are due to the Markov chain $\mathbb{F} - (W_1, Q_2^{[k]}, \mathbb{G}) - A_2^{[k]}, k = 1, 2$, which is proved as follows.

$$
\begin{aligned}
I(A_2^{[k]}; \mathbb{F} | W_1, Q_2^{[k]}, \mathbb{G}) &\leq I(A_2^{[k]}, S_2; \mathbb{F} | W_1, Q_2^{[k]}, \mathbb{G}) && (86)\\
&= I(S_2; \mathbb{F} | W_1, Q_2^{[k]}, \mathbb{G}) + I(A_2^{[k]}; \mathbb{F} | W_1, Q_2^{[k]}, \mathbb{G}, S_2) && (87)\\
&\stackrel{(8)}{=} I(S_2; \mathbb{F} | W_1, Q_2^{[k]}, \mathbb{G}) && (88)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \quad I(S_2, W_2; \mathbb{F}|W_1, Q_2^{[k]}, \mathbb{G}) && (89)\\
&= \quad I(W_2; \mathbb{F}|W_1, Q_2^{[k]}, \mathbb{G}) + I(S_2; \mathbb{F}|Q_2^{[k]}, \mathbb{G}, W_1, W_2) && (90)\\
&\stackrel{(4)}{\leq} \quad I(W_2; \mathbb{F}, W_1, Q_2^{[k]}, \mathbb{G}) && (91)\\
&\stackrel{(7)(6)}{=} \quad 0 && (92)
\end{aligned}
$$

On the other hand, from (70), we have

$$
\begin{aligned}
L \quad &\leq \quad I(W_2; Q_1^{[1]}, A_1^{[1]}|W_1, \mathbb{G}) + I(W_2; Q_2^{[1]}, A_2^{[1]}|W_1, \mathbb{G}) && (93)\\
&\stackrel{(11)}{=} \quad I(W_2; Q_1^{[2]}, A_1^{[2]}|W_1, \mathbb{G}) + I(W_2; Q_2^{[2]}, A_2^{[2]}|W_1, \mathbb{G}) && (94)\\
&\leq \quad I(W_2; Q_1^{[2]}, A_1^{[2]}, \mathbb{F}|W_1, \mathbb{G}) + I(W_2; Q_2^{[2]}, A_2^{[2]}, \mathbb{F}|W_1, \mathbb{G}) && (95)\\
&\stackrel{(6)}{=} \quad I(W_2; Q_1^{[2]}, A_1^{[2]}|W_1, \mathbb{F}, \mathbb{G}) + I(W_2; Q_2^{[2]}, A_2^{[2]}|W_1, \mathbb{F}, \mathbb{G}) && (96)\\
&\stackrel{(7)(8)(4)}{=} \quad H(A_1^{[2]}|W_1, \mathbb{F}, \mathbb{G}) + H(A_2^{[2]}|W_1, \mathbb{F}, \mathbb{G}) && (97)
\end{aligned}
$$

Combining (75), (81) and (97), we have shown that $H(A_2^{[2]}|W_1, \mathbb{F}, \mathbb{G}) = L/2$. The proof of (63) is complete.

Next we prove (64). On the one hand,

$$
H(A_2^{[2]}|W_1, A_2^{[1]}, \mathbb{F}, \mathbb{G}) \leq H(A_2^{[2]}|W_1, \mathbb{F}, \mathbb{G}) \stackrel{(63)}{=} L/2 \qquad (98)
$$

On the other hand, from sub-modularity of entropy functions we have

$$
\begin{aligned}
&H(A_2^{[2]}, A_2^{[1]}|W_1, \mathbb{F}, \mathbb{G})\\
&\quad \geq \quad -H(A_2^{[1]}, A_1^{[1]}|W_1, \mathbb{F}, \mathbb{G}) + H(A_1^{[1]}, A_2^{[2]}, A_2^{[1]}|W_1, \mathbb{F}, \mathbb{G}) + H(A_2^{[1]}|W_1, \mathbb{F}, \mathbb{G}) && (99)\\
&\quad \stackrel{(67)(13)(74)}{\geq} \quad -L/2 + H(A_1^{[1]}, A_2^{[2]}, A_2^{[1]}, W_2|W_1, \mathbb{F}, \mathbb{G}) + L/2 && (100)\\
&\quad \geq \quad H(W_2|W_1, \mathbb{F}, \mathbb{G}) \stackrel{(6)}{=} L && (101)\\
&\quad \Rightarrow \quad H(A_2^{[2]}|W_1, A_2^{[1]}, \mathbb{F}, \mathbb{G}) = H(A_2^{[2]}, A_2^{[1]}|W_1, \mathbb{F}, \mathbb{G}) - H(A_2^{[1]}|W_1, \mathbb{F}, \mathbb{G}) \stackrel{(74)}{\geq} L/2 && (102)
\end{aligned}
$$

Note that the second term of (100) follows from the assumption that $A_1^{[1]} = A_1^{[2]}$ so that from $A_1^{[1]}, A_2^{[2]}$, we can decode W_2 just as from $A_1^{[2]}, A_2^{[2]}$, we can decode W_2. Combining (98), (102), we have proved $H(A_2^{[2]}|W_1, A_2^{[1]}, \mathbb{F}, \mathbb{G}) = L/2$. Symmetrically, it follows that $H(A_2^{[2]}|W_2, A_2^{[1]}, \mathbb{F}, \mathbb{G}) = L/2$. Therefore, the desired inequality (64) is obtained.

■

From Lemma 3, we know that $I(A_2^{[1]}; A_2^{[2]}|W_1, \mathbb{F}, \mathbb{G}) = I(A_2^{[1]}; A_2^{[2]}|W_2, \mathbb{F}, \mathbb{G}) = 0$. Plugging in Ingleton's inequality [31] that holds for linear schemes but not for non-linear schemes, we have

$$
\begin{aligned}
I(A_2^{[1]}; A_2^{[2]}|\mathbb{F}, \mathbb{G}) \quad &\leq \quad I(A_2^{[1]}; A_2^{[2]}|W_1, \mathbb{F}, \mathbb{G}) + I(A_2^{[1]}; A_2^{[2]}|W_2, \mathbb{F}, \mathbb{G}) + \underbrace{I(W_1; W_2|\mathbb{F}, \mathbb{G})}_{=0, \text{ from } (6)}\\
&= \quad 0 && (103)\\
\Rightarrow H(A_2^{[1]}, A_2^{[2]}|\mathbb{F}, \mathbb{G}) \quad &= \quad H(A_2^{[1]}|\mathbb{F}, \mathbb{G}) + H(A_2^{[2]}|\mathbb{F}, \mathbb{G}) && (104)
\end{aligned}
$$

$$\begin{aligned}
&\overset{(61)}{=} && H(A_2^{[1]}|\mathbb{F},\mathbb{G}) + H(A_1^{[1]}|\mathbb{F},\mathbb{G}) && (105)\\
&\geq && H(A_1^{[1]}, A_2^{[1]}|\mathbb{F},\mathbb{G}) && (106)\\
&\overset{(13)}{=} && H(W_1, A_1^{[1]}, A_2^{[1]}|\mathbb{F},\mathbb{G}) && (107)\\
&= && H(W_1|\mathbb{F},\mathbb{G}) + H(A_1^{[1]}, A_2^{[1]}|W_1,\mathbb{F},\mathbb{G}) && (108)\\
&\overset{(6)}{\geq} && L + H(A_1^{[1]}|W_1,\mathbb{F},\mathbb{G}) \overset{(63)}{=} 3L/2 && (109)\\
\Rightarrow \alpha = H(S_2)/L &\geq && H(A_2^{[1]}, A_2^{[2]}|\mathbb{F},\mathbb{G})/L \geq 3/2 && (110)
\end{aligned}$$

5.1 Symmetrization

Theorem 3 *Consider the single-round PIR problem with $K = 2$ messages and $N = 2$ databases. Suppose we have a scheme described by $\bar{L}, \bar{W}_1, \bar{W}_2, \bar{S}_1, \bar{S}_2, \bar{Q}_{1:2}^{[1]}, \bar{Q}_{1:2}^{[2]}, \bar{A}_{1:2}^{[1]}, \bar{A}_{1:2}^{[2]}, \bar{\mathbb{F}}, \bar{\mathbb{G}}$. Then we can construct a symmetric PIR scheme, also for $K = N = 2$, described by $L, W_1, W_2, S_1, S_2, Q_{1:2}^{[1]}, Q_{1:2}^{[2]}, A_{1:2}^{[1]}, A_{1:2}^{[2]}, \mathbb{F}, \mathbb{G}$ such that*

$$\begin{aligned}
H(A_1^{[1]}|\mathbb{F},\mathbb{G}) &= H(A_2^{[1]}|\mathbb{F},\mathbb{G}) = H(A_2^{[2]}|\mathbb{F},\mathbb{G}) && (111)\\
H(S_1) &= H(S_2) && (112)\\
L &= 2\bar{L} && (113)
\end{aligned}$$

such that the symmetric PIR scheme has the same rate and storage overhead as the original PIR scheme.

Proof: Consider two independent implementations of the asymmetric PIR scheme. Let us use the 'bar' notation for the first implementation and the 'tilde' notation for the second implementation. In the first implementation, there are two messages $\bar{W}_1, \bar{W}_2$, each of length $\bar{L}$, two databases $\bar{\text{DB}}1$ and $\bar{\text{DB}}2$ which store $\bar{S}_1, \bar{S}_2$, respectively. In the second implementation, there are two messages $\tilde{W}_1, \tilde{W}_2$, each of length $\tilde{L} = \bar{L}$, two databases $\tilde{\text{DB}}2$ and $\tilde{\text{DB}}1$ which store $\tilde{S}_1, \tilde{S}_2$, respectively. Note the critical detail that the database indices are switched in the second implementation. The asymmetric PIR scheme specifies the queries for each implementation such that the user can privately retrieve an arbitrarily chosen message from each implementation.

The symmetric PIR scheme is created by combining the two implementations. In the combined scheme, there are two messages $W_1 = (\bar{W}_1, \tilde{W}_1)$ and $W_2 = (\bar{W}_2, \tilde{W}_2)$, each of length $L = 2\bar{L}$, two databases DB1 and DB2 which store $(\bar{S}_1, \tilde{S}_2)$ and $(\bar{S}_2, \tilde{S}_1)$, respectively. Retrieval works exactly as before. For example, if the user wishes to privately retrieve $W_1 = (\bar{W}_1, \tilde{W}_1)$, then it retrieves $\bar{W}_1$ exactly as in the first implementation, and $\tilde{W}_1$ exactly as in the second implementation.

Since the symmetric scheme is comprised of two independent implementations of the original PIR scheme, the message size, total download size, total storage size, are all doubled relative to the original PIR scheme. As a result the rate and storage overhead, which are normalized quantities, remain unchanged in the new scheme. Symmetry is achieved because each database from the original PIR scheme is equally represented within each database in the new PIR scheme.

Mathematically,

$$\begin{aligned}
W_1 = (\bar{W}_1, \tilde{W}_1), W_2 = (\bar{W}_2, \tilde{W}_2) && (114)\\
S_1 = (\bar{S}_1, \tilde{S}_2), S_2 = (\bar{S}_2, \tilde{S}_1) && (115)
\end{aligned}$$

$$\mathbb{F} = (\bar{\mathbb{F}}, \tilde{\mathbb{F}}), \mathbb{G} = (\bar{\mathbb{G}}, \tilde{\mathbb{G}}) \tag{116}$$

$$Q_n^{[k]} = (\bar{Q}_n^{[k]}, \tilde{Q}_{3-n}^{[k]}), n = 1, 2, k = 1, 2 \tag{117}$$

$$A_n^{[k]} = (\bar{A}_n^{[k]}, \tilde{A}_{3-n}^{[k]}) \tag{118}$$

where each random variable with a bar symbol is independent of and identically distributed with the same random variable with a tilde symbol. We are now ready to prove the first equality in (111).

$$\begin{aligned}
H(A_1^{[1]}|\mathbb{F}, \mathbb{G}) &= H(\bar{A}_1^{[1]}, \tilde{A}_2^{[1]}|\mathbb{F}, \mathbb{G}) && (119)\\
&= H(\bar{A}_1^{[1]}|\bar{\mathbb{F}}, \bar{\mathbb{G}}) + H(\tilde{A}_2^{[1]}|\tilde{\mathbb{F}}, \tilde{\mathbb{G}}) && (120)\\
&= H(\tilde{A}_1^{[1]}|\tilde{\mathbb{F}}, \tilde{\mathbb{G}}) + H(\bar{A}_2^{[1]}|\bar{\mathbb{F}}, \bar{\mathbb{G}}) && (121)\\
&= H(\bar{A}_2^{[1]}, \tilde{A}_1^{[1]}|\mathbb{F}, \mathbb{G}) && (122)\\
&= H(A_2^{[1]}|\mathbb{F}, \mathbb{G}) && (123)
\end{aligned}$$

where (120) and (122) follow from the fact that the two copies of the given scheme are independent and (121) is due to the property that the two copies are identically distributed. Consider the second equality in (111).

$$\begin{aligned}
H(A_2^{[1]}|\mathbb{F}, \mathbb{G}) &\overset{(7)}{=} H(A_2^{[1]}|Q_2^{[1]}, \mathbb{F}, \mathbb{G}) && (124)\\
&= H(A_2^{[1]}|Q_2^{[1]}, \mathbb{G}) && (125)\\
&\overset{(11)}{=} H(A_2^{[2]}|Q_2^{[2]}, \mathbb{G}) && (126)\\
&= H(A_2^{[2]}|Q_2^{[2]}, \mathbb{F}, \mathbb{G}) && (127)\\
&\overset{(7)}{=} H(A_2^{[2]}|\mathbb{F}, \mathbb{G}) && (128)
\end{aligned}$$

where (125) and (127) are due to the Markov chain $\mathbb{F} - (Q_2^{[k]}, \mathbb{G}) - A_2^{[k]}, k = 1, 2$, which is proved as follows.

$$\begin{aligned}
I(A_2^{[k]}; \mathbb{F}|Q_2^{[k]}, \mathbb{G}) &\leq I(A_2^{[k]}, S_2; \mathbb{F}|Q_2^{[k]}, \mathbb{G}) && (129)\\
&= I(S_2; \mathbb{F}|Q_2^{[k]}, \mathbb{G}) + I(A_2^{[k]}; \mathbb{F}|Q_2^{[k]}, \mathbb{G}, S_2) && (130)\\
&\overset{(8)}{=} I(S_2; \mathbb{F}|Q_2^{[k]}, \mathbb{G}) && (131)\\
&\leq I(S_2, W_1, W_2; \mathbb{F}|Q_2^{[k]}, \mathbb{G}) && (132)\\
&= I(W_1, W_2; \mathbb{F}|Q_2^{[k]}, \mathbb{G}) + I(S_2; \mathbb{F}|Q_2^{[k]}, \mathbb{G}, W_1, W_2) && (133)\\
&\overset{(4)}{\leq} I(W_1, W_2; \mathbb{F}, Q_2^{[k]}, \mathbb{G}) && (134)\\
&\overset{(7)(6)}{=} 0 && (135)
\end{aligned}$$

Finally, we prove (112).

$$\begin{aligned}
H(S_1) &= H(\bar{S}_1, \tilde{S}_2) && (136)\\
&= H(\bar{S}_1) + H(\tilde{S}_2) && (137)
\end{aligned}$$

$$
\begin{aligned}
&= H(\tilde{S}_1) + H(\bar{S}_2) && (138)\\
&= H(\bar{S}_2, \tilde{S}_1) && (139)\\
&= H(S_2) && (140)
\end{aligned}
$$

where (137) and (139) follow from the fact that the two copies of the given scheme are independent and (138) is due to the property that the two copies are identically distributed. ■

6 Conclusion

We showed that the capacity of MPIR is equal to the capacity of PIR, both with and without T-privacy constraints. Our result implies that there is no advantage in terms of capacity from multiround over single-round schemes, non-linear over linear schemes, or ϵ-error over zero-error schemes. We also offered a counterpoint to this pessimistic result by exploring optimal storage overhead instead of capacity. Specifically, we constructed a simple multiround, non-linear, ϵ-error PIR scheme that achieves a strictly smaller storage overhead than the best possible with any single-round, linear, zero-error PIR scheme. The simplicity of the scheme makes it an attractive point of reference for future work toward understanding the role of linear versus non-linear schemes, zero-error versus ϵ-error capacity, and single-round versus multiple round communications. Another interesting insight revealed by the scheme is the privacy benefit of reduced storage overhead. By not storing all the information at each database, and by optimally compressing the stored information, not only do we reduce the storage overhead, but also we enable privacy where it wouldn't hold otherwise.

References

[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, 1995, pp. 41–50.

[2] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private Information Retrieval," *Journal of the ACM (JACM)*, vol. 45, no. 6, pp. 965–981, 1998.

[3] R. Ostrovsky and W. E. Skeith III, "A Survey of Single-database Private Information Retrieval: Techniques and Applications," in *Public Key Cryptography–PKC 2007*. Springer, 2007, pp. 393–411.

[4] W. Gasarch, "A Survey on Private Information Retrieval," in *Bulletin of the EATCS*, 2004.

[5] S. Yekhanin, "Private Information Retrieval," *Communications of the ACM*, vol. 53, no. 4, pp. 68–73, 2010.

[6] H. Sun and S. A. Jafar, "The Capacity of Private Information Retrieval," *arXiv preprint arXiv:1602.09134*, 2016.

[7] ——, "Optimal Download Cost of Private Information Retrieval for Arbitrary Message Length," *arXiv preprint arXiv:1610.03048*, 2016.

[8] ——, "The Capacity of Robust Private Information Retrieval with Colluding Databases," *arXiv preprint arXiv:1605.00635*, 2016.

[9] ——, "The Capacity of Symmetric Private Information Retrieval," *arXiv preprint arXiv:1606.08828*, 2016.

[10] K. Banawan and S. Ulukus, "The Capacity of Private Information Retrieval from Coded Databases," *arXiv preprint arXiv:1609.08138*, 2016.

[11] Q. Wang and M. Skoglund, "Symmetric Private Information Retrieval For MDS Coded Distributed Storage," *arXiv preprint arXiv:1610.04530*, 2016.

[12] T. H. Chan, S.-W. Ho, and H. Yamamoto, "Private Information Retrieval for Coded Storage," *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, pp. 2842–2846, 2015.

[13] R. Dougherty, C. Freiling, and K. Zeger, "Insufficiency of linear coding in network information flow," *IEEE Transactions on Information Theory*, vol. 51, no. 8, pp. 2745 – 2759, Aug. 2005.

[14] T. H. Chan and A. Grant, "Dualities between entropy functions and network codes," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4470 – 4487, Oct. 2008.

[15] S. Rouayheb, A. Sprintson, and C. Georghiades, "On the Index Coding Problem and Its Relation to Network Coding and Matroid Theory," *IEEE Trans. on Inf. Theory*, vol. 56, no. 7, pp. 3187–3195, July 2010.

[16] A. Blasiak, R. Kleinberg, and E. Lubetzky, "Lexicographic products and the power of non-linear network coding," *ArXiv:1108.2489*, Aug. 2011. [Online]. Available: http://arxiv.org/abs/1108.2489

[17] M. Langberg and M. Effros, "Network Coding: Is zero error always possible?" in *49th Allerton Conference on Communication, Control and Computing.*, 2011, pp. 1478–1485.

[18] S. Jalali, M. Effros, and T. Ho, "On the impact of a single edge on the network coding capacity," in *Information Theory and Applications Workshop (ITA), 2011*. IEEE, 2011, pp. 1–5.

[19] O. Kosut and J. Kliewer, "On the relationship between edge removal and strong converses," in *Proceedings of International Symposium on Information Theory (ISIT)*, 2016.

[20] A. Beimel and Y. Ishai, "Information-theoretic private information retrieval: A unified construction," in *Automata, Languages and Programming*. Springer, 2001, pp. 912–926.

[21] A. Beimel, Y. Ishai, and E. Kushilevitz, "General constructions for information-theoretic private information retrieval," *Journal of Computer and System Sciences*, vol. 71, no. 2, pp. 213–247, 2005.

[22] N. Shah, K. Rashmi, and K. Ramchandran, "One Extra Bit of Download Ensures Perfectly Private Information Retrieval," in *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, 2014, pp. 856–860.

[23] A. Fazeli, A. Vardy, and E. Yaakobi, "Codes for distributed PIR with low storage overhead," in *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, 2015, pp. 2852–2856.

[24] R. Tajeddine and S. E. Rouayheb, "Private Information Retrieval from MDS Coded Data in Distributed Storage Systems," *arXiv preprint arXiv:1602.01458*, 2016.

[25] S. Rao and A. Vardy, "Lower Bound on the Redundancy of PIR Codes," *arXiv preprint arXiv:1605.01869*, 2016.

[26] S. Blackburn and T. Etzion, "PIR Array Codes with Optimal PIR Rate," *arXiv preprint arXiv:1607.00235*, 2016.

[27] T. E. Simon R. Blackburn and M. B. Paterson, "PIR schemes with small download complexity and low storage requirements," *arXiv preprint arXiv:1609.07027*, 2016.

[28] Y. Zhang, X. Wang, H. Wei, and G. Ge, "On private information retrieval array codes," *arXiv preprint arXiv:1609.09167*, 2016.

[29] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on information Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 2006.

[31] A. W. Ingleton, "Representation of matroids in combinatorial mathematics and its applications," *Combinatorial Mathematics and Its Applications*, vol. 44, pp. 149 – 167, Jul. 1971.